Generative Judge for Evaluating Alignment

Year: 2023 | Citations: 137 | Authors: Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao

---

Abstract

The rapid development of Large Language Models (LLMs) has substantially expanded the range of tasks they can address. In the field of Natural Language Processing (NLP), researchers have shifted their focus from conventional NLP tasks (e.g., sequence tagging and parsing) towards tasks that revolve around aligning with human needs (e.g., brainstorming and email writing). This shift in task distribution imposes new requirements on evaluating these aligned models regarding generality (i.e., assessing performance across diverse scenarios), flexibility (i.e., examining under different protocols), and interpretability (i.e., scrutinizing models with explanations). In this paper, we propose a generative judge with 13B parameters, Auto-J, designed to address these challenges. Our model is trained on user queries and LLM-generated responses under massive real-world scenarios and accommodates diverse evaluation protocols (e.g., pairwise response comparison and single-response evaluation) with well-structured natural language critiques. To demonstrate the efficacy of our approach, we construct a new testbed covering 58 different scenarios. Experimentally, Auto-J outperforms a series of strong competitors, including both open-source and closed-source models, by a large margin. We also provide detailed analysis and case studies to further reveal the potential of our method and make a variety of resources public at https://github.com/GAIR-NLP/auto-j.